The College Board Review

Intermediate Tests See First Trial

In this issue

May 1950 series runs ahead of June 1949 candidates

More than 16,000 candidates on May 20 took the examinations of the College Board. Last year, when the tests were held June 4, the number was about 12,500. Mr. Frank H. Bowles, Director of the Board, attributed the rise to three causes: increased use of Board tests by more colleges, the earlier testing date, and, to a lesser degree, the growing use of scores for the counseling of preliminary (junior-year) candidates.

College Ability Test to become regular feature of program

On May 13, almost exactly fifty years after the first College Board Tests were offered, the Intermediate Tests for College Students, consisting of a College Ability Test and five Proficiency Tests, were administered for the first time to 776 applicants for transfer from one college to another. At the first college entrance examinations half a century ago there were 973 candidates of whom 758 were seeking admission to either Columbia College or Barnard College. This year there will be over 75,000 candidates for the regular tests.

Beginning on August 9, 1950, the College Ability Test, a high ceiling measure of scholastic aptitude expressly designed for college transfer applicants, will become a regular feature of the Board's testing program. The Proficiency Tests in Humanities, Life Sciences, Mathematics, Physical Sciences, and Social Sciences will not be offered in 1950–51, except by special arrangement.

The CAT will be administered at the same centers and on five of the six testing dates already announced for the regular college entrance tests: August 9, 1950, December 2,

THE COLLEGE BOARD REVIEW
News and Research of the
College Entrance Examination Board
Published from time to time by the
College Entrance Examination Board
425 West 117th Street, New York 27, N. Y.
Director . . . . . Frank H. Bowles
Secretary . . . . . William C. Fels

1950, January 13, 1951, May 19, 1951, and August 15, 1951. Whether the CAT will be offered on March 10, 1951, the sixth date, has not been decided. The decision will depend upon demand, and upon whether the two programs can be administered simultaneously to the large number of March candidates without confusion at the centers or delay in the reporting of scores to the colleges.

The fee for the CAT will be $6.00.

Duke University admitted at April Board meeting

Duke University was admitted to membership at the April meeting of the Board. It becomes the one hundred and fifteenth member college, the sixteenth south of the Mason-Dixon line. Mrs. W. S. Persons, Director of Admissions of the Woman's College, and Mr. E. B. Weatherspoon, Director of Admissions, Trinity College and the College of Engineering, will represent the University.

Also elected was the American Association of Collegiate Registrars and Directors of Admission. The Association has designated Mr. Elwood C. Kastner, Registrar and Supervisor of Admissions at New York University, as its representative.

Fiftieth Anniversary meeting plans announced

The Fiftieth Anniversary meeting of the Board will be held at the Biltmore Hotel in New York on Tuesday, October 24, to be followed by a formal dinner at the University Club. Presidents Conant of Harvard and Eisenhower of Columbia, whose predecessors, Presidents Eliot and Butler, were instrumental in founding the Board, will be among the speakers. On Wednesday, October 25, also at the Biltmore, the Board will hold an invitation conference on admission to American colleges. President Arthur S. Adams of the University of New Hampshire, recently elected to the presidency of the American Council on Education, President Harold W. Stoke of Louisiana State University, Headmaster John Hallowell of Western Reserve Academy, and Professor Francis L. Bacon of the University of California, formerly Superintendent of Schools in Evanston, Illinois, are expected to speak.

"Irregular" English reports obsolete, discontinued

The practice of sending a special memorandum to accompany the score of a candidate who completed only one or two of the three questions or sections of the English Composition Test has been discontinued.

Now that the English Composition Test is no longer made up of three essay questions, almost no candidates (3 per cent in April 1949) fail to complete a substantial part of the test. There is now little justification for reporting this type of irregularity when no similar step is taken for the other achievement tests.

College choice rule to come up again in October

The much-discussed and often-changed "college choice" rule will be discussed and perhaps changed again at the Board's October meeting. The rule requires candidates to indicate their choice—or lack of choice—of college on the Board's application blank. This information is reported to the colleges.

Criticism of the present incarnation of the rule came to a head at the April meeting when it was called unwieldy, expensive, and not very useful for selecting students or estimating the number who may be expected to accept admission. The rule will be a special order on the October agenda.

Under the present rule, voted in October 1948, a candidate names as many as three colleges, indicating whether he ranks them 1–1–1, 1–1–2, 1–2–2, or 1–2–3. The old rule, before October 1948, required him to indicate colleges in 1–2–3 order. This was objected to on the grounds that it did not permit a true statement of preference, and the rule was changed.

Dissatisfaction with the new rule stems partly from the "asterisk problem." Special score reports marked with asterisks are sent to colleges for two classes of candidates: (1) those who, at the time of applying for examination, have not decided what colleges they will apply to; (2) those who ask, either on their original applications or later, that scores be reported to additional colleges. Colleges are unable to distinguish between these two classes of candidates. Although this is a clerical matter which could be corrected, to do so would be surprisingly costly. It would involve individual checking and changing of thousands of records. It would also delay score reporting materially.

Dates, Tests, Fees: 1950-51

EXAMINATION DATES

August 9, 1950
December 2, 1950
January 13, 1951
March 10, 1951
May 19, 1951
August 15, 1951

EXAMINATION PROGRAMS*

Morning Program

Scholastic Aptitude Test
(Verbal Section)
(Mathematical Section)

Afternoon Program

Achievement Tests

(a maximum of three afternoon tests)

English Composition
Social Studies
French Reading
German Reading
Greek Reading
(March only)
Italian Reading
(March only)
Latin Reading
Spanish Reading
Biology
Chemistry
Physics
Intermediate Mathematics
Advanced Mathematics

Aptitude Tests

Pre-Engineering Science Comprehension
Spatial Relations

EXAMINATION FEES

Morning Program and Afternoon Program	$12.00
Morning Program only	6.00
Afternoon Program only	8.00

* For information concerning the College Ability Test, see p. 137.

Progress report—Committee on English Testing

Frank D. Ashburn

Mr. Frank D. Ashburn is Headmaster of Brooks School, North Andover, Massachusetts, and President of the New England Association of Colleges and Secondary Schools. He is Chairman of the Subcommittees on Achievement Tests and English Testing and until recently was Chairman of the Committee on Examinations of the College Entrance Examination Board.

For some time an increasing number of loyal supporters and constant users of Board tests has been concerned with what has seemed a lack in the program. They have admired the present tests for their predictive reliability, their economy in time and money, their flexibility. They have at no time thought the objective tests should be given up.

Their basic concern was that when the essay type test was abandoned there ceased to be any testing of certain crucially important educational values, and they dreaded that unless these values could be re-emphasized, or possibly even rediscovered, the effect on American education would be most unfortunate.

First discussions led to the conclusion that while no test has received more competent study, the existing English test is the least satisfactory offered by the Board, and that there is no ground for hope that it can be materially improved because of time limitations imposed. It became clear that numerous schools and colleges sensed the lack of some educational ingredient. The Committee therefore considered what this ingredient might be. They decided to their own satisfaction that it was identifiable, definable, larger than English, although present in English and possibly best measured through English.

They felt that there can be a product of education which leads to power to deal cogently and lucidly with ideas synthesized from reading, study, and experience both academic and practical. This product, scarcely measured at all at present and in danger of being lost altogether, is one, be it repeated, which is not peculiar to English or subject to departmentalization, but is the essence of the entire educational process. Its development is the hall mark of a good school and its possession the best indication of good college material and, indeed, of good education at any level.

The Committee therefore shifted its emphasis from literature to what it called tentatively Comprehensive Composition and later decided to call General Composition. It developed a trial run test, three hours in length, of essay type, with three questions set, one from each of the three fields of literature, science, and social studies, from which the candidate was to select two, on each of which he would write for an hour and a half. Three forms were used.

SCHEME OF MARKING

A revolutionary scheme of marking was set up. Only three symbols were designated: O for outstanding, A for adequate, and I for inadequate, although by combining any two the total gradations were increased to five (with such intriguing results as outstandingly-adequate or adequately-outstanding, or adequately-inadequate; there was one happy suggestion that OI, outstandingly-inadequate, conveyed useful information). Instead of a single mark, four ratings were given, one for content, one for organization, one for mechanics, and one for style. The reading was to be distributed between English teachers and instructors in the other fields involved.

Through the extraordinary cooperation of heads and departments, the trial run was given in February to more than 1000 pupils in some twenty schools and colleges. These were picked

to give variety in size and kind: large and small, boys and girls, public and private. Some college freshmen also took part. The papers were given and corrected at the schools by the students' own teachers. The interest and effort of the teachers were remarkable and gratifying.

FIRST TEST UNSATISFACTORY

From this trial run interesting information was gained, which may be summarized briefly. The actual tests were not satisfactory because badly drawn. It was not satisfactory to have a separate test in science. The results in social studies were scarcely better, because the ordinary high school pupil has not had enough work in either of these fields. Those wishing some form of essay testing were in the ratio 35–1. An overwhelming majority approved the basic idea of the test, though many were critical of details. There was evidence that exact ratings are not necessary and that there are gains in not attempting them. There was a striking reduction in reading time when they were not required and when a reader read for one and not for all the four categories. The actual average reading time, although there was wide variation, astonishingly did not exceed ten minutes. Another surprising thing was that there was little difference in the ratings for the 11th and 12th grades. And finally, there were painfully few O's in any school, and department after department accepted responsibility for this, rather than laying it on the pupils, and stated that giving the tests had had a wholesome effect on the faculty.

With this material in mind the Committee, after long discussion, made the following decisions, again abbreviated here. For the present there should be no further attempt to offer separate tests in science and social studies, but a single question should be framed so as to allow the student with a strong bent in either of these fields to use knowledge and show power. Two hours of writing were held ample, but at least forty-five minutes were held necessary for preliminary definition and illustration. It was agreed that separate ratings rather than a single score should be retained, but that the essence of the problem was careful instruction for candidates and readers. Since the objective is the classification of students into groups rather than exact ratings, it is not necessary to rely too heavily on such things as the probability curve.

The earliest practicable date at which a General Composition Test could be offered as part of the regular Board program would be in May–June, 1951. Another trial run, this time a true pretest, should be held next autumn, preferably with another selected sampling of schools, but with additional copies of tests available for any schools who wished to experiment, even though not in the sampling. The first regular test would be thought of as primarily for 11th graders, though obviously available for 12th graders not using it for final college admission. The tentative plan is that for a time, at any rate, the schools will be allowed to administer the tests in their own rooms, without sending candidates to regular Board centers, and that they should not be held down to a single date, but rather allowed a reasonable period, on any day of which the tests could be given. This would permit a flexibility which would not interfere with the present program and would give a participating school a choice of using the SAT in the morning and the General Composition in the afternoon, or of using SAT in the morning, achievement tests in the afternoon, and the General Composition Test at some convenient time near the end of the academic year.

TESTS MAY BE AVAILABLE

The Committee's hope is that for a brief period the new tests can be made available without extra cost, but that within not more than three years they will become self-supporting. For reasons of administration, control, and analysis it would be necessary to limit the number of "official" tests in the first program to perhaps 2500, probably on a first-come-first-served basis, but an unlimited number could be made available to interested schools for their own administration and correction.

Questioning the Questions

John Landis and Frances Swineford

This is the fourth of the series of articles surveying the Board's testing and research programs. Mr. John Landis is head of the Science and Engineering Section of the Test Development Department and Dr. Frances Swineford is head of the Test Analysis Section of the Statistical Analysis Department of the Educational Testing Service.

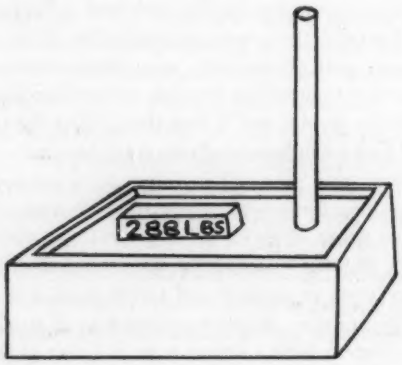
Each year, in each of the major academic fields, examiners of the College Entrance Examination Board are confronted with a heterogeneous mass of questions (largely of their own making) from which they must sort out the ingredients of a good, new test. Just how do these examiners—committees of school and college teachers—decide which questions shall be used? As Mary Turnbull indicated in her article on "Building the Board's Tests" in the last issue of the *Review*, they work closely with the test construction specialists of Educational Testing Service in making their selections. But they do not depend on subjective judgment alone. Every effort is made to try out, or pretest, each question, or item, before it is used in a final form, the chief purpose being to weed out the items which do not differentiate among students at the college-entrance level.

Groups of students of the same age, academic background, and scholastic aptitude as the candidates who will take the Board tests are chosen for these tryouts. Usually a sample of about 400 students is employed. These are the "guinea pigs" to whom the diet of raw but likely-looking items selected by the various committees is fed. Their answers to each question are carefully analyzed and the discriminating power and the difficulty of the question are evaluated on the basis of these answers. We shall discuss the analysis phase of test construction at greater length later.

Each pretest is administered under conditions which conform to Board regulations, and care is taken to insure high motivation. When the testing situation is controlled in this way, and the pretest results are analyzed properly, the test-makers are in a good position to select a well-balanced final form. Both sides of the picture are represented: on the one hand, the choices and requirements of the examiners; on the other, the understanding and abilities of the candidates.

Thus, not only five examiners, but 400 of a candidate's peers aid in the construction of each test instrument which evaluates his fitness for college.

To illustrate the fallibility of utilizing examiner judgments only in setting up a Board test, consider the following two items, both of which were added to the Board Physics files in 1944. Both were rated as excellent questions by the Physics Examiners, yet pretesting proved one to be a dud. Can you pick out the good item?



1. An upright pipe connects with the tank of a hydrostatic balance illustrated in the diagram. The area of the platform is 4 square feet, that of the cross-section of the pipe

1 square inch. If a 288-pound load is placed on the moving platform

(a) water rises in the pipe until the weight of water in the pipe equals that of the load (b) the water would rise ½ as far if the pipe were twice the area of cross-section (c) the platform and water in the pipe move equal distances (d) this device depends on Bernoulli's Principle (e) the water in the pipe rises a little over 1 foot

Key—(e)

2. A glass tube is bent so as to form a tall U tube. Water is poured into one side and oil is poured into the other so that the boundary between them is just in the middle of the bend at the bottom of the U tube. If the water column is 10 centimeters high and the oil column is 12.5 centimeters high, we may conclude that

(a) oil is heavier than water (b) the specific gravity of the oil is 1.25 (c) the density of the oil is 0.80 gm. per cc. (d) the specific gravity of the water is 1.25 (e) the density of the oil is 0.25 gm. per cc.

Key—(c)

Though a diagram usually increases the validity of a Physics item, it didn't help the first item in this case. The second is the good item.

How do we know? Our analysis shows that on the first item too high a proportion of the better students fell for the plausible wrong answers (a) and (b) perhaps through hurried reading. And of course we want to exclude from the test those items which in effect trap the good students (the very ones we want to make high scores) but not the poorer candidates. Yet it is highly improbable that the performance of students on this particular item would have been predicted accurately by any group of physics teachers without the aid of the pretest.

To aid your understanding of how test construction specialists go about interpreting the results of pretests to select the good items from the bad, consider the following simplified test situation. Suppose that you, as a classroom teacher, have administered a multiple-choice test consisting of twenty five-choice items to a class of twenty-four pupils. You mark the papers and find that the scores range from a high of 19 to a low of 9. Arranging the papers in order of score, you find that the scores are distributed according to Table 1. Now you are ready to start an item analysis.

TABLE 1

Score	Number of Papers
19	1
18	3
17	2
16	3
15	5
14	4
13	1
12	2
11	2
10	—
9	1
Total	24

If you were to separate the papers into four groups of six each, with the highest scores in the first group, the next highest in the second, and so on, you would expect the largest number of correct answers to any single item to occur in the highest group and the smallest number in the lowest group. Instead of comparing all four groups, let us compare just the highest and lowest groups. Table 2 is the beginning of the tabulation you might devise for summarizing your results.

TABLE 2

Item	1	2	3	4	5	6	7
Number in high group answering correctly	6	5	2	4	6	6	3
Number in low group answering correctly	2	1	2	3	6	4	5

Most of your items are probably good ones, and they will really help to pick out the able pupils from the poor pupils. Numbers 1, 2, and 6 are examples. Number 6 looks easier than numbers 1 and 2, for more of the low group answered it correctly; nevertheless, all three items discriminate between the high and low groups. Number 3 is difficult even for the high group and does not discriminate at all. Number 4 does not have so large a difference between the two

groups as we like to find. Number 5 is too easy, for everybody in both groups answered it correctly. Although it may be a good question to put in a test of basic points which must be mastered by everyone in your class, it would not belong in a College Board test, for example, because it is of no help in selecting the best students. Item number 7 is an example of a very poor item. More members of your low group got it right than of your high group. Perhaps there was some ambiguity in the wording of the item, or possibly one or more of the "distractors" (the incorrect choices) was poorly expressed.

A similar analysis of each distractor would give you even more information about the excellence of your items. If no one marks a given distractor, for example, that particular item is, in effect, a four-choice item instead of the five-choice item that you intended it to be. Moreover, a good distractor is one which is selected by more of the low group than of the high group. Not just one, but all the distractors should be good ones.

If you think we have outlined a pretty big task for you, consider now how much bigger it is for us, even though we do have a group of obliging machines to do part of the work. Instead of one test of twenty items, we analyze the Scholastic Aptitude Test, which may have two hundred items, and from thirteen to fifteen achievement tests, each of which contains more than one hundred items. Since we want results upon which we can rely with confidence, we generally use as many as four hundred or five hundred papers instead of the twenty-four that you used.

STUDY MORE GROUPS

Finally, we do not stop with just four ability groups and study the highest and lowest, as you did, but we use more than twenty groups. From them we obtain a single number, the "biserial correlation," which expresses the discriminating power of the item. The biserial correlation can be as high as 1.00 or as low as -1.00. If it is high, say .60 or .70, the item resembles your items 1 and 2. If it is zero, the item is like your items 3 and 5, and if it is negative, the item is like your item 7.

For the more statistically-minded reader, we have reproduced an item analysis sheet for a discarded Chemistry item. This item was chosen because it yields some interesting results, which will be explained in detail at the end of the article. The discussion immediately following the sheet pertains to the more technical aspects of item analysis.

PROBLEM OF ANALYSIS

The primary problem of item analysis can be restated succinctly: What is the relationship between ability to answer a particular item and grasp of the subject being tested? The only practicable criterion of a student's subject-matter understanding available to us is his total score on the test of which the item is a part. Therefore, we ask the question: Did the high scorers on the test answer this particular item correctly for the most part—and did the low scorers on the test answer it incorrectly? If so, the item, by the total-test standard, is a good one. If all the people who scored above a certain level gave the right answer to this item and all who scored below this level gave the wrong answer, then the biserial correlation for the item (r_{bis} on the item sheet) would be +1.00. If the reverse were true, i.e., all the high scorers missed this item and all the low scorers selected the right answer, then r_{bis} would be −1.00. Normally, of course, r_{bis} does not approach these limits.

In general, we like our tests to be sufficiently homogeneous that the majority of the biserial correlations are over .30. Many are .60 or higher. Items with low correlations are either rewritten and tried again or discarded altogether.

Let us now look in detail at the illustrative item-analysis sheet to see what information it gives us. First, at the top of the page is the identification. From this we know that we are examining item number 80 of a chemistry test given to a sample of 430 students. The item itself

is reproduced in the middle of the page. It is a five-choice item, for which the correct response is the third choice, indicated by the heavy rules above and below "Response Code" number 3.

In the N_r column are the numbers of students selecting the various responses coded in the first column. Two people omitted the item entirely, 116 marked the correct response (number 3),

Card Number | ITEM 0 8 0

TEST CHEM
FORM V C B A 1
BASE N. 430

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
ITEM ANALYSIS

Date Tabulated 11 47 Operator Number 8

Response Code	N_r	Σy	Mean	LOW N_1	N_2	N_3	N_4	N_5	HIGH N_6
0	2	15		2					
1	35	439	12.5	3	7	10	11	2	2
2	10	125	12.5		3	3	2	2	
3	116	1527	13.164	19	14	19	21	26	17
4	23	272	11.8	5	3	3	6	6	
5	244	3242	13.3	33	37	30	58	51	35

80. The valence of an element is equal to the number of

(1) its atoms combined with 1 atom of hydrogen (2) atoms of oxygen combined with 1 atom of this element (3) atoms of hydrogen combined with 1 atom of this element (4) protons in the nucleus of its atoms (5) electrons in the valence shell of its atoms

	N_t	Σy_t	My_t	Σy_t^2	σy_t
Total Tried (t)	430	5620	13.070	79896	3.870

$\frac{N+}{N_t}$ = p: .270 $\quad My_t + \sigma y_t x =$ Δ: 15.4 $\quad \left(\frac{M+ - My_t}{\sigma y_t}\right)\left(\frac{p}{z}\right) =$ r_{bis}: .02

Computed by_____
Checked by_____

NOTES:-

1. x, based on p, is the distance from the mean along the baseline of the normal curve, in terms of unit standard deviation. If p is less than .5, x is positive; if p is more than .5, x is negative.
2. Compute means for responses made by ten or more candidates when r_{bis} is less than .40.
3. Compute M+ and My_t to three decimal places; means of all other responses, to one.
4. Record r_{bis} to two places; Δ to one.
5. Carry all other computations to three decimal places.

FORM 573 SM 1PT 8-49

4555

and a greater number, 244, were drawn to the incorrect fifth response.

In order to be able to compute the biserial correlation (r_{bis}), already discussed, we must have the average score on the total test for the 116 students who answered correctly. The tabulating-machine portion of this work has been simplified by transmuting the original scores to a scale for which the mean is 13 and the standard deviation is 4. Only integral values of the new scores are used, and this rounding process results in a mean and standard deviation which differ slightly from the desired figures. Near the bottom of the item-analysis sheet you will see that they are, respectively, 13.070 and 3.870.

The column headed Σ_y gives the sum of the scores on the test as a whole for each group of students. From these sums and the N_r values we can compute the average total test score for each group. If the item is measuring whatever the test as a whole measures, the average should be high for those selecting the correct response and low for all the other groups. Here, it is interesting to note in the column headed "Mean" that the group who selected the fifth response has the highest average—an indication that there is something wrong with the option. This response is not only the most popular one, but it is attracting too many of the most able students.

FURTHER BREAKDOWN

One further breakdown of the scores is made in order to give still more information. The total group is divided into six subgroups on the basis of the total test score. These groups are identified by N_1 (the lowest-scoring group), $N_2, \ldots, N_6$ (the highest-scoring students) at the right of the page. The numbers in the row which corresponds to the correct response should increase as one moves from N_1 to N_6 (i.e., a larger number of the better students should answer correctly) and those for all the wrong responses should decrease. In this item, however, responses 3 and 5 have much the same distribution, whereas responses 1, 2, and 4 do not exhibit the desired characteristic of wrong responses. Sometimes only one distractor fails to show the decline in numbers from N_1 to N_6, and a slight revision in the wording of that distractor may suffice to reclaim the item. Unfortunately this is not the case with the item we are considering.

TWO IMPORTANT STATISTICS

At the bottom of the page are two important statistics: r_{bis}, which has already been discussed, and p. The statistic, p, is the proportion of all students trying the item who answered it correctly. Thus p is a measure of the difficulty of the item. In the example a little over one-fourth of the group was successful, showing that this item is difficult for the present group. A test which discriminates well between the members of the group must yield scores throughout a wide range of values, with no undue bunching at any particular point. Thus the test must not be too easy or too difficult. By having on file the difficulty values of a great many items, members of our test development department are able to assemble a test which is accurately pitched as to difficulty.

As we have said, item analysis is useful not only to aid in the selection of items which are suited to a particular group of students, but also to indicate how certain of the poor items can be revised for future pretesting. For an example of the latter use, we refer you again to our Chemistry sample. The figures on the item-analysis sheet focus attention immediately on choices (3) and (5). Choice (3) is regarded as the correct answer. But now that the pretest results raise doubts, we ask, "Is it really correct?" Then the light begins to dawn. Choice (3) is an incomplete definition of valence because it covers only those elements which will combine with hydrogen; that is, mainly the elements with negative valence. For better results, it should have been worded: "(3) atoms of hydrogen *or atoms of chlorine* combined with 1 atom of this element."

This partially explains why some of the better students went on and chose (5) as the correct answer. Though (5) neglects the possibility of

multiple valence, it is almost as good a definition of positive valence as (3) is of negative valence. In addition, the wording of (5) probably attracts the modern student to it. It mentions both "electrons" and "valence shell"—two resounding terms that smack of current terminology.

We can improve this item in either of two ways. To make an easy question out of it, we can use (3) as reworded and modify (5) to read "orbital electrons of its atom." To make a more difficult question out of it, we can set the choices up as follows:

(1) its atoms combined with 1 atom of hydrogen
(2) its atoms combined with 2 atoms of oxygen
(3) free electrons in the outer shell of its atom
(4) electron-proton bonds in its atom
(5) electron-pair bonds which its atom shares with other atoms

Here (5) is the correct answer. Choices (2) and (4) were modified because only ten people chose (2) and twenty-three people with a low mean score chose (4).

Even at this stage we cannot say whether or not the item will be a good one. We must pretest it again before we put it into a final test. Who can say what the analysis will show?

Slow—but How Sure?

W. G. Mollenkopf

Dr. W. G. Mollenkopf is a Research Associate of the Educational Testing Service.

As students are filing out of an examination room, how often is the remark heard, "If I'd only had a little more time, I could really have shown what I can do!"

Many teachers have felt concerned over the fairness of a speeded test to those whom they regard as "slow but sure." That well over half of those who take the SAT finish neither the Verbal nor the Mathematical section is revealed by the Annual Reports of the Director of the College Board for 1947 and 1948. Since the slow-working student is at an obvious disadvantage on a speeded test, one may well ask, "Is the SAT fair to the 'slow but sure'?"

Before going further we should pause and ponder a bit about that word "fair." Fair in what way? Clearly the questioner has in mind the use made of the scores. Hence he must mean fair in terms of selection for admission to college. Suppose that some applicants get low SAT scores because they work slowly, even though they possess good ability to handle the test material when given plenty of time. To the extent that these students would do well in college despite low SAT scores, the test is unfair, for these students might well have difficulty in being admitted to college in view of their low scores.

This suggests that the best way to determine whether SAT is fair to the "slow but sure" would be through an experimental study in which materials of the types contained in SAT were administered to college applicants at varying rates of speed. Students would be followed-up, and their college grades related to their scores on the different tests, in order to find out the predictiveness of the tests given at various speeds. This kind of experiment would reveal whether there are some slow but competent individuals who do well on a test only when given

adequate time, and who *also* do well in college. It would be valuable for the College Board to support such a study.

SPEEDING OF SAT

On what basis, then, was the present rate of speeding of material decided upon for SAT? The answers are perhaps two: first, to insure good testing conditions, and second, to secure a wide range of scores. Test supervisors report that if most students finish before the time is up, this leads to restlessness and noise in the testing room, sometimes to such an extent that those who have not yet finished are adversely affected. The second point is tied up with the basic notion of measurement. Our yardstick needs length and an adequate number of units along its scale so that we can provide fairly precise indices of the relative abilities of individuals.

Naturally both of these points are important. However, even though I would not want to belittle the need for good order while tests are being given, I feel confident that our test supervisors could manage matters even when a high proportion of candidates finished the tests, if it were demonstrated that better measurement of applicants could be gained by such a change.

The second point, then, would seem to strike closer to the heart of the matter. It is quite true generally that the spread of scores obtained for a test is wider when the time given the students is so short that only a few finish than it is when most have time to try all of the items. Now a test can be successful as a measuring device only when different persons get different scores. The basic question, then, is whether the speeded test measures the same thing as the unspeeded test. That is, does the speeding of the test make the scores reflect something different from what they indicate when the students are given plenty of time?

Some experimental evidence from a research study is available on this point. The data indicate that for one type of material used in the SAT, practically the same rankings of students are obtained when the test is speeded as when it is unspeeded. However, for another type of material, added time does make a difference. The slow and accurate student comes out near the top when given plenty of time but suffers when the time limit is short. To present the findings best, let us consider the research in some detail.

In making a study of the effects on item-analysis data of changing item placement and test timing, a Verbal Antonyms test and a Mathematics Aptitude test were constructed at the Educational Testing Service. These tests contained items (i.e., test questions) similar to those appearing in the College Board Scholastic Aptitude Test. Two forms of each test were prepared, each pair containing the same items but in different orders.

The cooperation of the Scarsdale, New York, High School was requested and received. This high school was a particularly appropriate place for the administration of the tests because it provided groups of appropriate size for making statistical analyses worth-while, because the student body was largely college preparatory, and because the students at the school were quite familiar with the procedures used in carrying out objective testing and could be expected to adapt readily to the unusual requirements of the study.

Few public high schools possess similar characteristics. The author wishes to express both his own personal thanks and the gratitude of the College Entrance Examination Board and the Educational Testing Service for the splendid cooperation received from the Scarsdale High School faculty and students.

STUDY AT SCARSDALE

After the purposes and the requirements of the study were outlined, arrangements for the testing were made by Mr. Lester W. Nelson, the Principal of the Scarsdale High School. Each staff member concerned with the testing was briefed concerning the purposes and the procedures involved. Furthermore, a notice was sent to each member of the junior and senior high-school classes, indicating that a research study

was to be carried out at the School which would aid in the construction of improved tests for the College Entrance Examination Board. The students were told that they would take two different tests, one on each testing day, which would involve different kinds of material and which would be given under different testing conditions. They were informed that the results would have no bearing on their school records. The importance of the study and the need for their cooperation were stressed in statements to the students.

Four comparable groups of students were made up out of the entire junior-senior student body by assigning to a particular group every fourth person whose name appeared on alphabetical lists arranged by grade and sex.

The Verbal and the Mathematics tests were administered at separate testing sessions to these groups. In each case, the students first worked under a time limit so fixed that few had sufficient time to try all of the items in the test. (This we shall call the *speed* condition.) Pencils were then changed from red to blue and the students allowed sufficient added time for practically everyone to attempt every question in the entire test. (This we shall call the *power* condition.) For the first phase, the students were instructed to work is if they were taking the usual speeded test, that is, to do the best they could in the time which was available to them. After the change of pencils, the students were then told both to attempt items which they previously had been unable to reach due to lack of time and also to go back to any item previously attempted and to change responses if they so desired. A change of response was indicated by means of an X marked through the answer to the item; the color of this X showed when the change in response had occurred. The actual testing went off quite smoothly, with the students cooperating splendidly.

Next we determined the number right both under the short time limit and under the longer time limit, and compared these scores, to find out whether the added period of time actually did benefit certain students. Of course, we expected that the added time would increase the scores of most students; the question we really raised was whether the added period of time served to change the relative standings of many students in the group. (If it did, then this would indicate that speed and power scores might not be used interchangeably.)

Table 1 presents some statistical data on the tests. The correlations between the two types of score for the Verbal tests were clearly very high,

TABLE 1

Means, Standard Deviations, and Correlations between Speed and Power Scores

Test Form	N	Speed Score Mean	Speed Score Std. Dev.	Power Score Mean	Power Score Std. Dev.	Speed-Power Correlation
Verbal 1	93	27.9	17.4	40.9	19.2	.94
Verbal 2	96	27.9	14.3	42.4	16.9	.91
Math. 1	100	9.8	4.8	18.8	7.8	.81
Math. 2	96	11.0	6.5	20.7	8.2	.82

whereas those for the Mathematics tests were substantial, though somewhat lower than those for the Verbal tests. Even with correlations of this magnitude, however, we realized that *some* individuals might have benefited greatly from the added period of time. A further—and important—point we had to keep clear about as we analyzed the data was that some shifts in relative standing are practically always to be expected when two sets of scores are obtained for the same person. Only with perfectly correlated measures would the standings be identical. With these points in mind, then, we examined the scatter plots of number right under speed versus number right under power to see whether there were a number of striking changes in relative standings. These plots are Figures 1–4 on the next page.

First let us look at Figures 1 and 2. It is clear that for the Mathematics tests the scatter plots may be aptly described as being somewhat fan-shaped. Evidently changes in relative standing occurred with considerably greater frequency through the middle range of scores than they did at either extreme of the range of scores. A

few of the more striking changes in relative standing have been identified by letters in the figures. Student A had a score of 8 points on Mathematics Form 1 under the speed condition, and a power score of 28. His speed score had a rank of 63 from the top, whereas his power score had a rank of 11.5. For Mathematics Form 2, student B had a score of 7 points under the speed condition and a power score of 30. His speed score ranked 67.5 from the top whereas his power score had a rank of 17.5. These two cases showed perhaps the most dramatic changes in relative standing between the two conditions of test administration, but many other changes in relative standing also occurred. For example, student C had a score of 13 points on Form 2

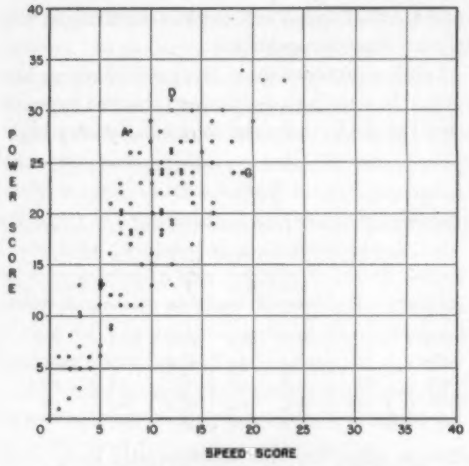


Figure 1. Scatter plot of speed and power scores on Form 1 of the Mathematics Aptitude test. (N = 100)

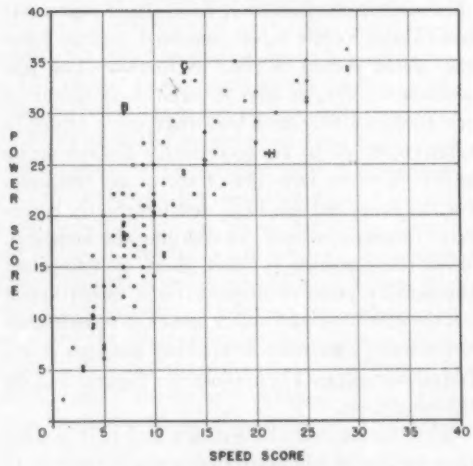


Figure 2. Scatter plot of speed and power scores on Form 2 of the Mathematics Aptitude test. (N=96)

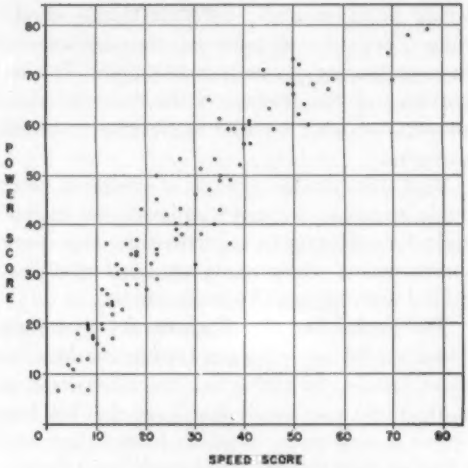


Figure 3. Scatter plot of speed and power scores on Form 1 of the Verbal Aptitude test. (N = 93)

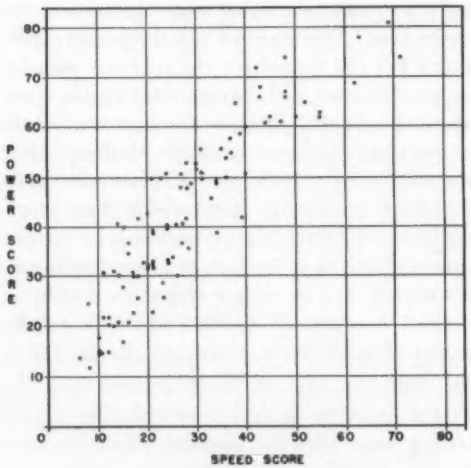


Figure 4. Scatter plot of speed and power scores on Form 2 of the Verbal Aptitude test. (N = 96)

of the Mathematics test under the speed condition, and this score had a rank of 27.5, whereas his power score of 34 points ranked 5th among the 96 power scores. On Form 1 of the examination, student D had a speed score of 12 points and a power score of 31 points. His speed score had a rank of 31.5, whereas his power score ranked 4th among the 100 power scores.

DECREASES IN STANDING

Naturally these increases in relative standing must be accompanied by some decreases. For example, consider the scores of student G. (See Figure 1.) His score of 19 had a rank of 4.5 among all the speed scores; so when the test is taken under a stringent time limit and the score is the number of questions answered correctly, he is rated as one of the few best students. Among the power scores, however, his power score of 24 had a rank of 29.5, and thus his power score is not included in the uppermost quarter of these scores. Similarly the speed score of student H ranked 10th highest, whereas his power score had a rank of 26.5. (See Figure 2.)

Another way in which the scatter plot may be examined is to answer the question, "For what range of speed scores were quite high power scores obtained?" Let us consider as high scores those which are more than a full standard deviation above the mean. For Mathematics Form 1, persons whose scores were thus "high" under the power condition (i.e., above 26.6) had speed scores ranging from 8 to 23 points—in other words, from slightly below the mean of the speed scores to the highest speed score obtained. For Form 2, "high" power scores (i.e., above 28.9) were obtained by persons whose speed scores ranged from 7 points to 29 points—that is to say, from more than half a standard deviation below the mean of the speed scores to the highest obtained speed score.

Let us turn now to the scatter plots for the Verbal Antonyms tests, Figures 3 and 4. It is at once apparent that these plots do *not* have the fan-shaped appearance noticed in the Mathematics scatter plots. Here there is the typical narrow elliptical shape usually characteristic of the very high correlations that were obtained. While it would have been possible for the plot to be broader for high power scores than for low power scores even with such a high correlation, this did not, in fact, happen. No dramatic changes of relative standing like those cited for the Mathematics tests can be given for the Verbal tests. In Figures 3 and 4 it is evident that practically all of the points fall in a narrow band of constant width. Those persons with the high speed scores all have high power scores. No one has an outstanding power score without having a distinctly better-than-average speed score as well. What differences in standing do occur with the Verbal tests are probably best regarded as due to random error—that is, as the sort of thing which can always be expected when two sets of measurements on the same persons are taken and the results compared.

CHANGES WITHIN GROUP

The data which we have just cited indicate that it is indeed true that some individuals make marked changes in their relative standings within the group when they are given added time during which to work on a test, especially a test of mathematics aptitude. Added time seems much more likely to allow a student to change his relative standing in the group when the type of item involved in the test is one requiring problem solving. When the type of item is one in which success on the item is largely a matter of recognition, as in the Verbal Antonyms test, the added time seems to have little if any effect on relative standing in the group.

In any event, the purpose for which the test is being given must be taken into account in interpreting the meaning of these results. If the purpose of the test is the selection of students for admission to college, that type of test should be given which either has the highest validity for predicting college success by itself or which when combined with other predictors adds most to the predictiveness of the combination.

Officers and Committees of the College Board

OFFICERS

Elected Officers

Chairman: President Katharine E. McBride, Bryn Mawr College
Vice-Chairman: Provost Samuel T. Arnold, Brown University
Chairman of the Executive Committee: Professor Emeritus George W. Mullins, Barnard College
Custodians:
Dr. Claude M. Fuess, Chestnut Hill, Mass., *Chief Custodian*
President James H. Case, Jr., Washington and Jefferson College
Vice-President Archibald MacIntosh, Haverford College

Appointed Officers

Director and Treasurer: Mr. Frank H. Bowles
Secretary: Mr. William C. Fels

COMMITTEES

Executive Committee

Professor Emeritus George W. Mullins, Barnard College, *Chairman* (ex officio)
Provost Samuel T. Arnold, Brown University (ex officio)
Mr. Frank D. Ashburn, Brooks School
Professor Francis L. Bacon, University of California
President Everett N. Case, Colgate University
Dean Margaret T. Corwin, New Jersey College for Women, Rutgers University
Miss Rosamond Cross, Baldwin School
Dr. Richard M. Gummere, Harvard University
Dean Radcliffe Heermance, Princeton University
Dr. Lemuel R. Johnston, Clifford J. Scott High School, East Orange, N. J.
Rev. Howard Kenna, University of Notre Dame
Dean Frank R. Kille, Carleton College
President Katharine E. McBride, Bryn Mawr College (ex officio)
Professor Edward S. Noyes, Yale University
Vice-President E. Kenneth Smiley, Lehigh University

Committee on Examinations

Dr. William H. Cornog, Central High School, Philadelphia, Pa., *Chairman*
Dr. Finla G. Crawford, Syracuse University
Miss Frances D. Dugan, Winsor School
Dean Wilma Kerby-Miller, Radcliffe College
Dean Millicent C. McIntosh, Barnard College
Dean Ernest C. Marriner, Colby College
Mr. Wilson Parkhill, Collegiate School

Committee on Research and Development

President Leonard Carmichael, Tufts College, *Chairman*
Dean Henry S. Dyer, Harvard University
Dean Sherwood R. Mercer, Muhlenberg College
Mr. Lester W. Nelson, High School, Scarsdale, N. Y.
President Rosemary Park, Connecticut College
Mr. William G. Saltonstall, Phillips Exeter Academy
Professor John M. Stalnaker, Illinois Institute of Technology

Committee on Finance

President Roswell G. Ham, Mount Holyoke College, *Chairman*
Mr. W. Emerson Gentzler, Columbia University
Mr. LeRoy E. Kimball, New York University
Dean C. Scott Porter, Amherst College

Committee on Audit

President Roswell G. Ham, Mount Holyoke College, *Chairman*
Dean C. Scott Porter, Amherst College

Committee on Nominations

President Roswell G. Ham, Mount Holyoke College, *Chairman*
Professor John M. Daniels, Carnegie Institute of Technology
Vice-President Archibald MacIntosh, Haverford College
Miss Catherine Rich, Catholic University of America
Mr. E. Laurence Springer, Pingry School